Comparing RR-Interval-Based and Whole-Signal-Based Machine Learning Models for Atrial Fibrillation Detection from Single-lead Electrocardiograms

Zixuan Ding[1], Jonathan Mant[1], James Brimicombe[1], Tommaso Bucci[2], Benjamin JR Buckley[2], Peter Calvert[2], Wern Yew Ding[2], Andrew Dymond[1], Gregory Y H Lip[2,3], Riccardo Proietti[2], Kate Williams[1], Elena Punskaya[1], Peter H Charlton[1], on behalf of the SAFER Investigators

[1] University of Cambridge, Cambridge, United Kingdom
[2] Liverpool Centre for Cardiovascular Science at University of Liverpool, Liverpool John Moores University and Liverpool Heart & Chest Hospital, Liverpool, United Kingdom
[3] Aalborg University, Aalborg, Denmark

Abstract

The aim of this study was to compare the performance of machine learning models to detect atrial fibrillation (AF) from single-lead ECGs which use either RR-intervals alone, or the entire ECG signal. Experiments were conducted using single-lead, 30-second ECG signals acquired using handheld ECG recorders from two datasets: the Computing in Cardiology (CinC) 2017 dataset (public), and the Screening for Atrial Fibrillation with ECG to Reduce Stroke (SAFER) dataset (private). The models assessed in this study were: two models which used the whole ECG signal, both of which were top-performing models from the 2017 PhysioNet / CinC Challenge; and two RR-interval-based models - a state-of-the-art model and a novel model which detects AF from a 2D representation of the differences between RR intervals. The models had AUROCs of 0.93 - 0.99. The AUPRCs varied more widely, from 0.64-0.94. The novel RR-interval-based AF detection model achieved an AUPRC of 0.94 on the CinC 2017 dataset, outperforming the state-of-the-art RR-interval-based model (0.88) and the entire-signal-based models (0.68 and 0.64). This experiment demonstrated that AF detection models utilizing only RR intervals could achieve comparable performance to those utilizing the entire ECG signal.

1. Introduction

Atrial Fibrillation (AF) is an irregular heart rhythm that increases the risk of having strokes by four to five times. AF is diagnosed from ECGs exhibiting no discernible repeating P waves and irregular RR intervals [1]. Models have been developed for detecting AF by extracting statistical and morphological features for use in machine-learning classifiers, as well as deep-learning neural networks that utilize the entire signal as the input; however, the potential of using RR intervals alone has been less thoroughly explored. Single-lead ECG devices are more susceptible to noise, potentially contaminating the P-wave section of the ECG which also places importance on investigating models which use RR intervals alone.

This experiment aimed to assess and compare the performance of state-of-the-art machine learning models for AF detection which take either RR-intervals alone, or the whole ECG signal, as inputs. The assessment was performed on single-lead ECGs collected using handheld devices in non-clinical environments. Additionally, a novel RR-interval-based approach was introduced, assessed and compared with the state-of-the-art approaches.

2. Methods

2.1. ECG Datasets and Pre-processing

The models were trained and tested separately on each of two single-lead ECG datasets: the public 2017 Computing in Cardiology dataset [2] (CinC 2017) and the private Screening for Atrial Fibrillation with ECG to Reduce Stroke (SAFER) dataset [3]. ECGs in both datasets were acquired using handheld ECG recorders, providing signals analogous to lead I ECGs. The characteristics of the datasets used in this experiment are presented in Table 1.

2.1.1. The CinC 2017 Challenge Dataset

The CinC 2017 challenge dataset [2] was collected using AliveCor's ECG device from participants who had purchased one of three generations of the device. All ECGs in the dataset are manually categorized as: 'Normal', 'AF', 'Other Rhythm', or 'Noisy'.

ECGs labelled 'Noisy' or with duration $<$30 s were ex-

cluded from this experiment to fulfil the European Society of Cardiology's minimum duration requirement [1]. This resulted in 7,655 out of 8,828 ECGs being included.

The publicly available dataset has designated training and validation sets, whilst the testing set is concealed. Therefore, the publicly available training set was randomly divided into new training and validation subsets, maintaining a ratio of 0.85 to 0.15 across all labels, and the publicly available validation set was used for testing. The original 'Normal' and 'Others' labels were combined into a single 'non-AF' class, while the original 'AF' class was retained.

The dataset was already pre-processed using a bandpass filter (0.5-40Hz), and normalized to "± 5 mV dynamic range" [2], so no further preprocessing was performed.

Table 1: The number of samples in each subset of both datasets.

Data	Subset	AF	non-AF	Total
CinC	Training	551	5,754	6,305
2017	Validation	97	1,016	1,113
	Testing	40	197	237
SAFER	Training	1,414	18,319	19,733
	Validation	279	4,179	4,458
	Testing	544	4,241	4,785

2.1.2. The SAFER Dataset

This study used anonymous data collected in the SAFER (Screening for Atrial Fibrillation to Reduce Stroke) Randomised Controlled Trial (ISRCTN 72104369) [3], collected from participants aged 70 or over using a Zenicor ECG device. Participants were asked to record an ECG four times a day for three weeks. ECGs were analysed by the Cardiolund algorithm [4], and any ECGs showing possible signs of AF were sent for manual review (*i.e.* which the algorithm classed as 'Irregular sequence', 'Fast regular heart rate', 'Fast episode', or 'Poor Quality'). Manual review was performed by two cardiologist teams, each reviewing half of the participants, and labelling a subset of ECGs. A diagnosis was provided for each participant, with AF diagnosed if one or more ECGs were labelled as AF. In this study we only used data from participants diagnosed as 'AF' or 'non-AF', and excluded those diagnosed as 'cannot exclude AF' or 'other significant arrhythmia'.

Participants were partitioned into training (70%), validation (15%), and test sets (15%), ensuring the ratio of each participant-level label was preserved in all subsets. The following ECGs were included: (i) all labelled ECGs from AF participants; for all non-AF participants, (ii) two ECGs randomly selected among all cardiologist-labelled 'non-AF' ECGs, one with 'Irregular Sequence' tag (where available), and one without an 'Irregular Sequence' tag (where available); and (iii) two ECGs which were not labelled by cardiologists were randomly selected, one with an 'Irregular Sequence' tag (where available), and one without an 'Irregular Sequence' tag (where available).

All signals from the SAFER database underwent Butterworth bandpass filtering between 0.5 and 40 Hz. The bandpassed ECGs were normalized so that the had an interquartile range of -0.5 to 0.5. In addition, when using the data with whole-signal-based models, which were adapted from CinC 2017 challenge entries, SAFER ECGs were resampled to a frequency of 300 Hz to match the original frequency of the challenge database.

2.2. Models

Two models that took whole ECG signals as inputs (Sec. 2.2.1), and two models that took RR intervals as inputs (Sec. 2.2.2) were evaluated. All models outputted a value between 0 and 1 indicating the likelihood of AF.

2.2.1. Whole-signal-based models

The whole-signal-based models used in this study consisted of one feature-based model [5] and one deep-learning model [6]. These were selected amongst the CinC 2017 Challenge entries with the top score [2]. They were retrained for binary AF/non-AF classification on both datasets.

The feature-based model [5] first extracted 188 features from the input ECG signal, including the mean, median of RR intervals, corrected QT interval (QTc), QR, QRS widths, Shannon, Tsallis and Renyi entropy using the author's submission to the challenge. An AdaBoost Regressor took in an input vector of the 188 features. The number of estimators in the regressor was set to 10.

The deep-learning model [6, 7] processes ECG segments, which were extracted from the original ECG recordings using a fixed window size of 10 seconds, while the stride (the size of the step moving on for each segment) varied according to the label in the training set (0.28 seconds for AF signals and 1.67 seconds for non-AF signals). The stride when segmenting the validation and test subsets was set to the sliding window for AF signals (*i.e.* 0.28 seconds). The final diagnosis of the 30-second segment was made from the majority diagnosis of the 10-second windows.

2.2.2. RR-interval-based models

RR intervals were derived from ECGs by computing the time difference between R peaks detected using the 'nk' beat detector in the Neurokit2 package [8].

A state-of-the-art RR-interval-based model was selected [9]. It was originally designed for diagnosing AF from continuous ECG measurements, and comprised two distinct components: firstly, a residual network taking a series of 60 RR intervals as an input and outputting a likelihood of AF between 0 and 1; and secondly a recurrent network which retained information from previous periods, and outputted an AF likelihood. For this study the model was adapted to accept 40 RR intervals as the input, and its second component was excluded. The RR intervals were padded with zeros if there were less than 40 RR intervals.

In addition, we developed a novel RR-interval-based model which takes a 2D representation of the differences between successive RR intervals, denoted dRR, as an input. This 2D input was an image similar to a Lorenz plot [10], with two modifications. Whilst a Lorenz plot is a scatter plot of the current dRR (dRR[i]) versus the previous dRR (dRR[i-1]), we also included differences between not only consecutive RR intervals, but also RR intervals separated by one, two, and three RR intervals. The plot axes were both restricted to -600ms to 600ms. Then, each axis was divided into 30 units, and the number of points falling into each square was calculated and plotted as a density plot. Figure 1a and 1c illustrate the plotting of dRR pairs, while Figure 1b and 1d display the corresponding density plot, where a brighter square indicates that it contained more points than a darker one. The neural network used to process this representation was a simplified version of Mobilenet-V2 [11], comprising only seven residual blocks. It is believed that our dRR representation was much simpler than real-life coloured images, for which the network was originally designed.

2.2.3. Training Process

All models were trained using the same training dataset. The whole-signal deep-learning model [6] and the RR-interval-based state-of-the-art model [9] underwent training using a batch size of 32 and a 0.001 learning rate, whereas the novel model was trained with a smaller batch size of 8 and a 0.0005 learning rate. For each model, from amongst the sets of model parameters generated after each training epoch, an optimal set was selected as the set with the minimum difference between the Area Under the Precision-Recall Curve statistics on the training and validation subsets.

2.3. Evaluating Model Performance

Sensitivity, Specificity, and Positive Predictive Value (PPV) were computed and plotted using the Receiver Operating Curve (ROC) and Precision-Recall (P-R) curve. Additionally, the Area Under the Curve (AUC) for all curves was calculated.

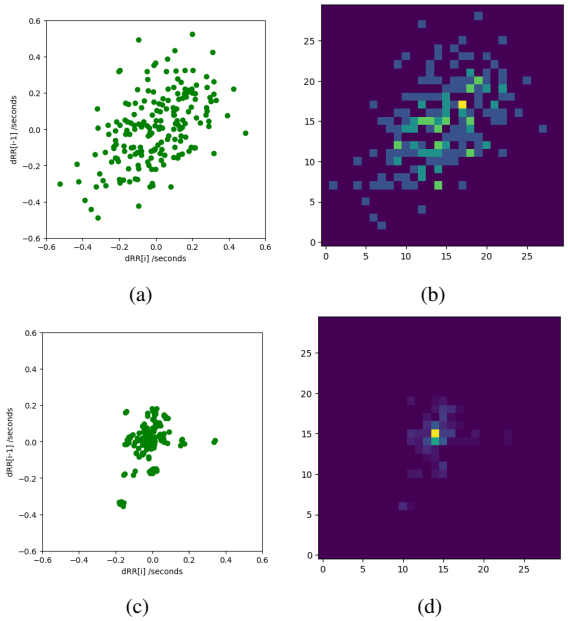
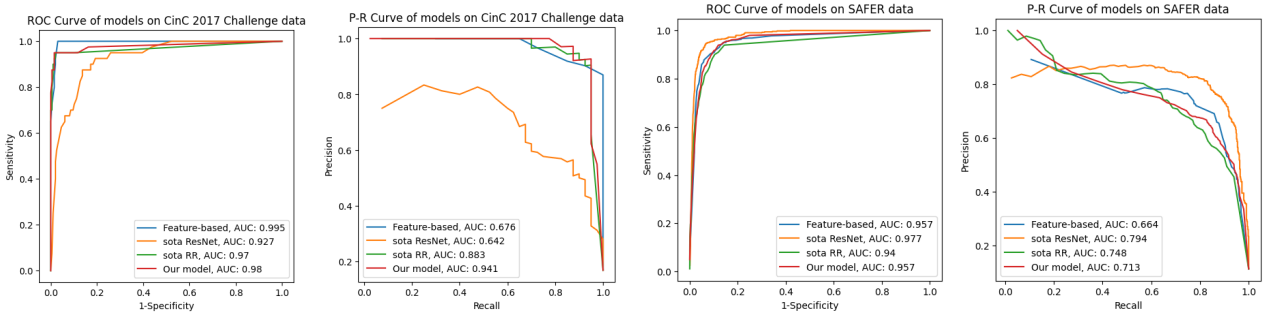


Figure 1: The 2D representation of differences between successive RR intervals used as an input to our proposed model: (a) and (c) show modified Lorenz plots for SAFER AF and non-AF ECGs, and (b) and (d) show their density representations, which were used as model inputs.

3. Results

The performance of the models is illustrated in Fig.2. For the CinC 2017 dataset, the AUROCs of the state-of-the-art [9] and proposed RR-interval-based models were similar to that of the whole-signal feature-based [5] model (AUROCs: 0.97, and 0.98, 0.995 respectively, Fig. 2a), whereas the AUROC for the whole-signal ResNet model [6, 7] was lower (0.927). This was likely because the ResNet model [6, 7] has more parameters than the others and was thus more susceptible to overfitting, especially considering the small size of the CinC 2017 dataset [2]. This was also illustrated in the P-R curve in Fig. 2b (AUPRCs: 0.883, 0.941 and 0.676 compared to 0.642).

For the SAFER dataset, the Receiver-Operating Curves of the models aligned closely (AUROCs: 0.94, 0.957, 0.957, and 0.977, Figure 2c), but the Precision-Recall curves exhibited significant variability (AUPRCs: 0.748, 0.713, 0.664, and 0.794, Figure 2d). This was because the SAFER dataset was more imbalanced than the CinC challenge dataset [2], which resulted in more false positives, and therefore lower precision, despite similar sensitivity and specificity. The P-R AUC for the whole-signal ResNet model (0.794) [6, 7] was slightly higher than that of the RR-interval-based models [9] (0.748 and 0.713) and the whole-signal feature-based model [5] (0.664).



(a) (b) (c) (d)

Figure 2: Results of model testing: (a) and (b) show Receiver Operating Characteristic (ROC) and Precision-Recall (P-R) curves for the CinC 2017 Challenge datasets; (c) and (d) show ROC and P-R curves for the SAFER dataset. Area Under the Curve (AUC) statistics are shown for each model: whole-signal 'Feature-based' [5] and 'sota ResNet' [6, 7] models; and RR-interval-based 'sota RR' [9] and 'our model' models.

4. Discussion and Conclusion

The main finding of this study was that RR-interval-based AF detection models could achieve comparable performance to those utilizing the whole ECG signal.

To the best of our knowledge all submissions to the CinC 2017 challenge [2] utilized the entire ECG signal as input. Thus, this study advances our understanding of using RR intervals alone for AF detection. Future work should consider additional performance metrics such as the Net Reclassification Index to assess the potential clinical utility.

Acknowledgment

This study is funded by a W.D. Armstong Trust Fund studentship awarded to ZD; and by the British Heart Foundation [FS/20/20/34626]. The SAFER Trial was funded by the National Institute for Health and Care Research (NIHR) [Programme Grant for Applied Research (RP-PG-0217-20007). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. For open access, the author(s) has applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising. ChatGPT (OpenAI, San Francisco, CA, USA) was used for language editing.

References

[1] Hindricks G, Potpara T, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). European Heart Journal 2021; 42(5):373–498.

[2] Clifford G, Liu C, et al. AF Classification from a Short Single Lead ECG Recording: the Physionet Computing in Cardiology Challenge 2017. 2017 Computing in Cardiology Conference. 2017; .

[3] Williams K, Modi RN, et al. Cluster randomised controlled trial of screening for atrial fibrillation in people aged 70 years and over to reduce stroke: protocol for the pilot study for the safer trial. BMJ Open 2022;12(9):e065066.

[4] Svennberg E, Stridh M, et al. Safe automatic one-lead electrocardiogram analysis in screening for atrial fibrillation. EP Europace 2017;19(9):1449–1453.

[5] Datta S, Puri C, et al. Identifying Normal, AF and other Abnormal ECG Rhythms using a Cascaded Binary Classifier. 2017 Computing in Cardiology Conference. 2017; .

[6] Hong S, Wu M, et al. ENCASE: an ENsemble ClASsifiEr for ECG Classification Using Expert Features and Deep Neural Networks. 2017 Computing in Cardiology Conference. 2017; .

[7] Hong S, Xu Y, et al. HOLMES: Health OnLine Model Ensemble Serving for Deep Learning Models in Intensive Care Units. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020; 1614–1624.

[8] Makowski D, Pham T, et al. Neurokit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods 2021;53(4):1689–1696.

[9] Biton S, Aldhafeeri M, et al. Generalizable and robust deep learning algorithm for atrial fibrillation diagnosis across geography, ages and sexes. npj Digital Medicine 2023; 6(11):1–10.

[10] Sarkar S, Ritscher D, Mehra R. A detector for a chronic implantable atrial tachyarrhythmia monitor. IEEE transactions on bio medical engineering 2008;55(3):1219–1224.

[11] Sandler M, Howard A, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks. CoRR 2018;abs/1801.04381.

Address for correspondence:

Zixuan Ding, zd295@cam.ac.uk

Strangeways Research Laboratory, 2 Worts' Causeway, Cambridge, CB1 8RN, United Kingdom